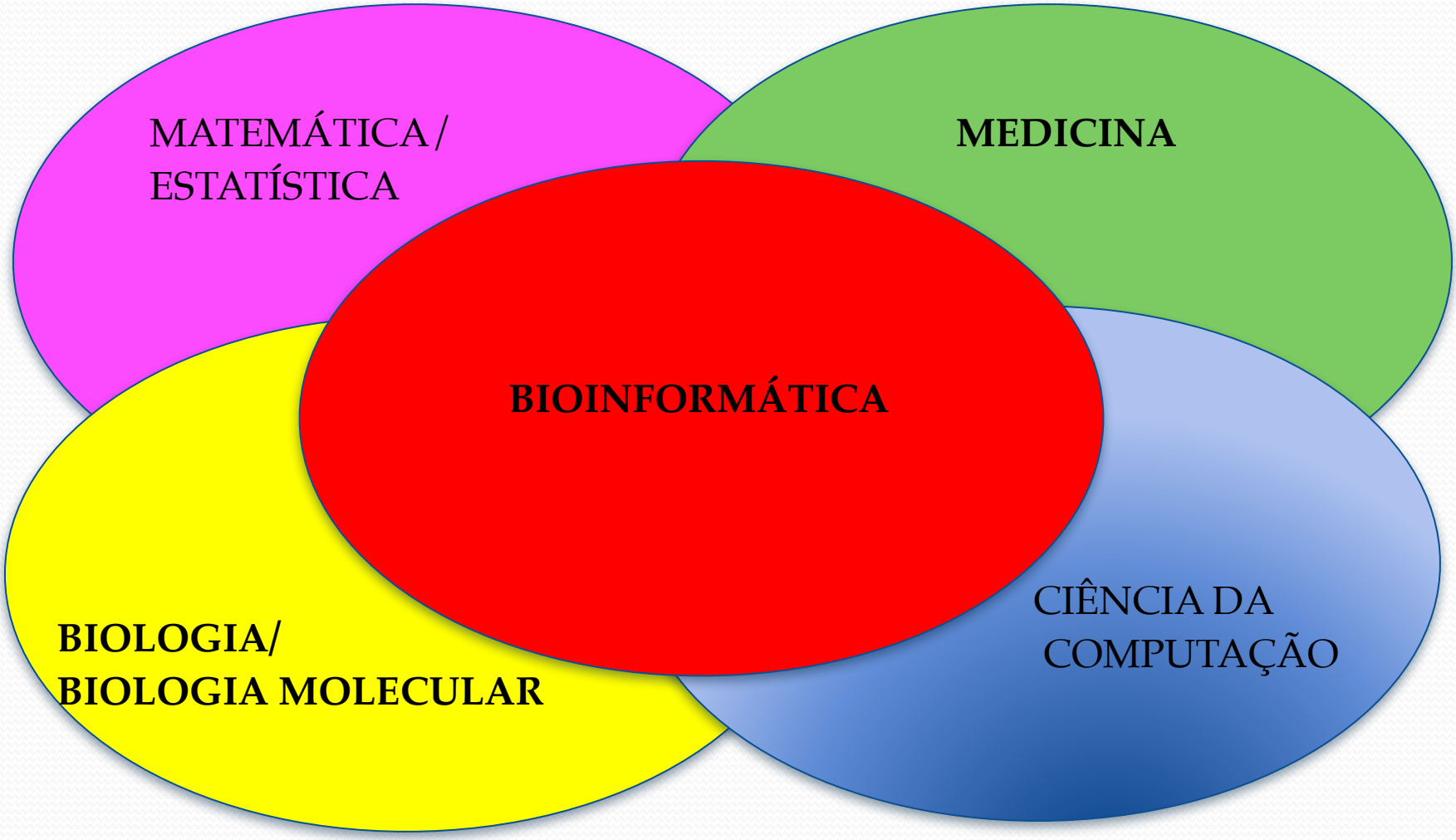
Bioinformática: que bicho é este?

Alan Mitchell Durham

Departamento de Ciência da Computação da USP

aland@usp.br

Bioinformática: que bicho é este?



Áreas que utilizam Bionformática

- Agronegócio
  - Melhoramento animal e vegetal
  - Combate a pragas
- Medicina/Veterinára
  - Caracterizaçào de patógenos
  - Diagnóstico preciso de doenças infecto-contagiosas
  - Desenvolvimento de vacinas
  - Diagnóstico precoce de doenças, incluindo as congênitas
  - Desenvolvimento de novos remédios
  - Análise de imagens diagnósticas
- Biologia em Geral
  - sequenciamento de genomas
  - Compreensão dos mecanismos de funcionamento da célula
  - Análise da arvore da vida
  - Classificação de microorganismos

Mas antes precisamos saber...

- COMO FUNCIONAM OS SERES VIVOS?
- (o dogma básico da Biologia Molecular)

Mas antes precisamos saber...

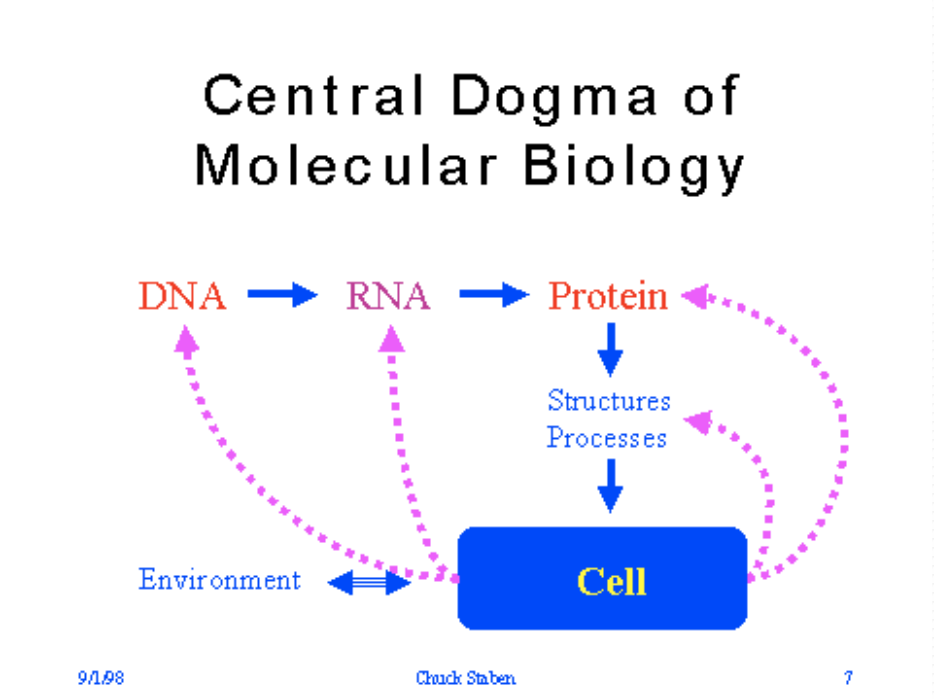
- COMO FUNCIONAM OS SERES VIVOS?
- (o dogma básico da Biologia Molecular)
- As proteínas são constituintes fundamentais de nosso corpo
  - Função estrutural
  - Controle do metabolismo
- Nossos cromossomos contém a informação essencial para formação do ser
  - Clonagem!
- Cromossomos contém o código para produção das proteínas

Dos Cromossomos às Proteínas (para computeiros)

Dos Cromossomos às Proteínas (para ~~computeiros~~ Cientistas da computação)

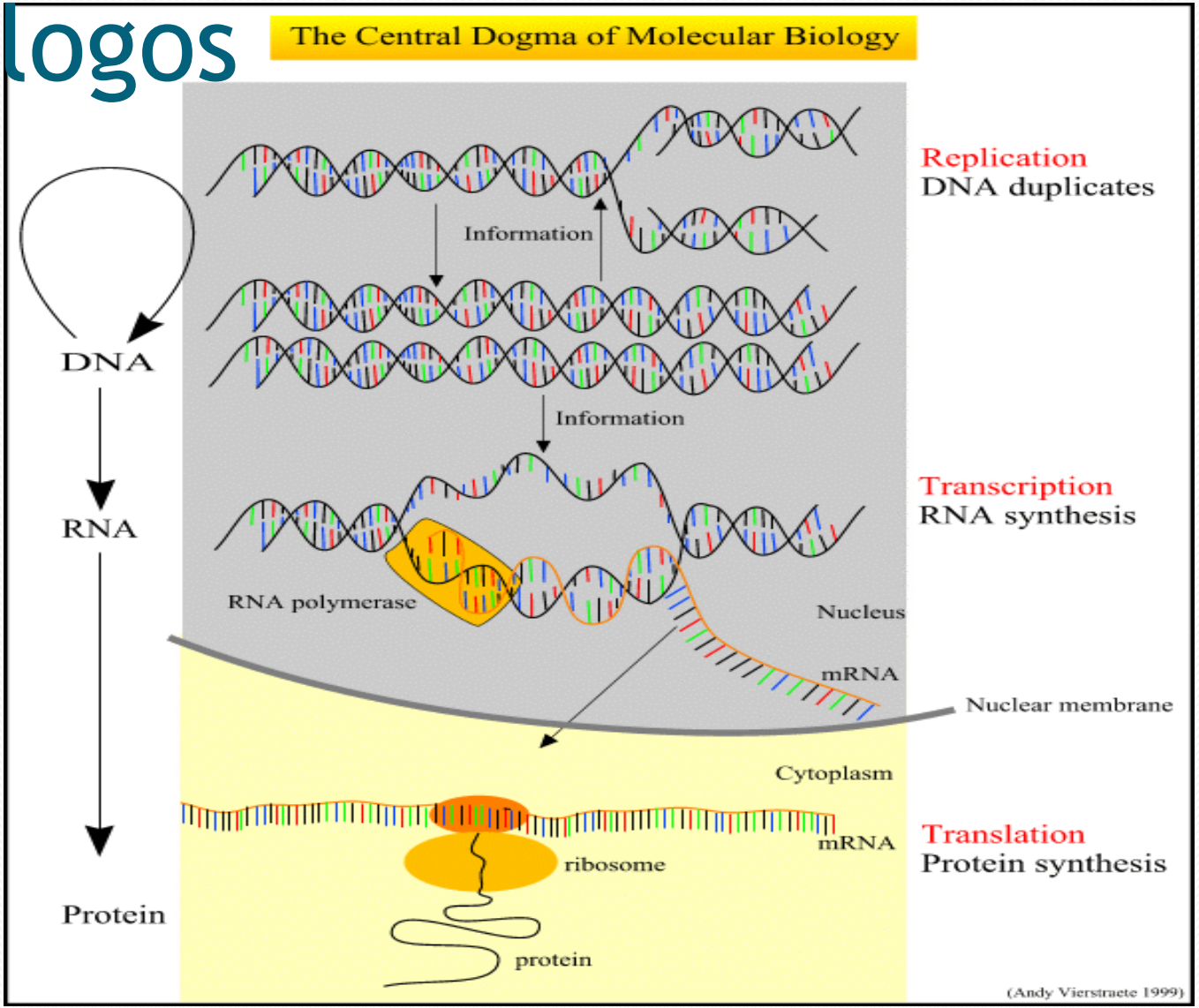
Fonte: http://www.uky.edu/Classes/BIO/520/BIO520WWW/Lecture2/lect2/sld007.ht

Dos Cromossomos às Proteínas (para ~~computeiros~~ Cientistas da computação)

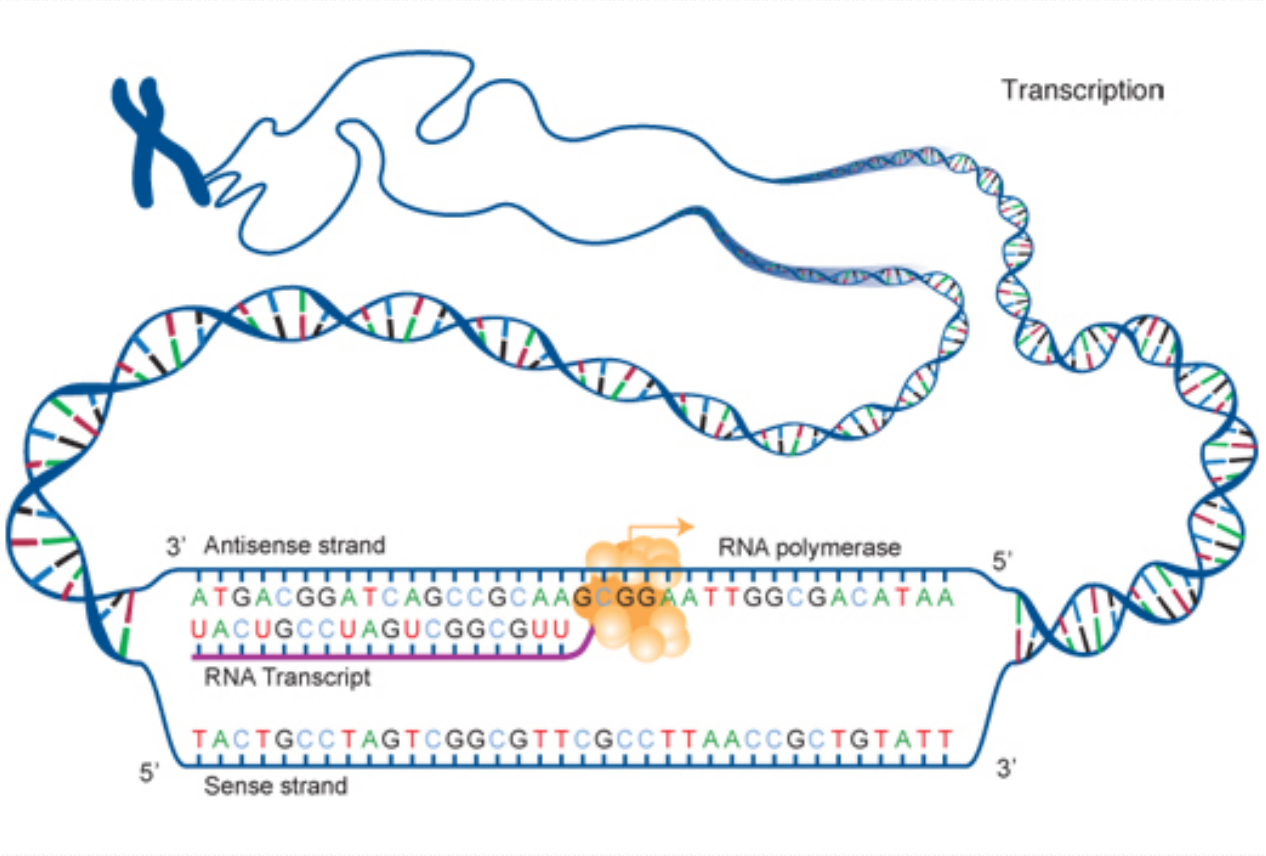


Fonte: http://www.uky.edu/Classes/BIO/520/BIO520WWW/Lecture2/lect2/sld007.ht

Dos Cromossomos às Proteínas: para Biólogos

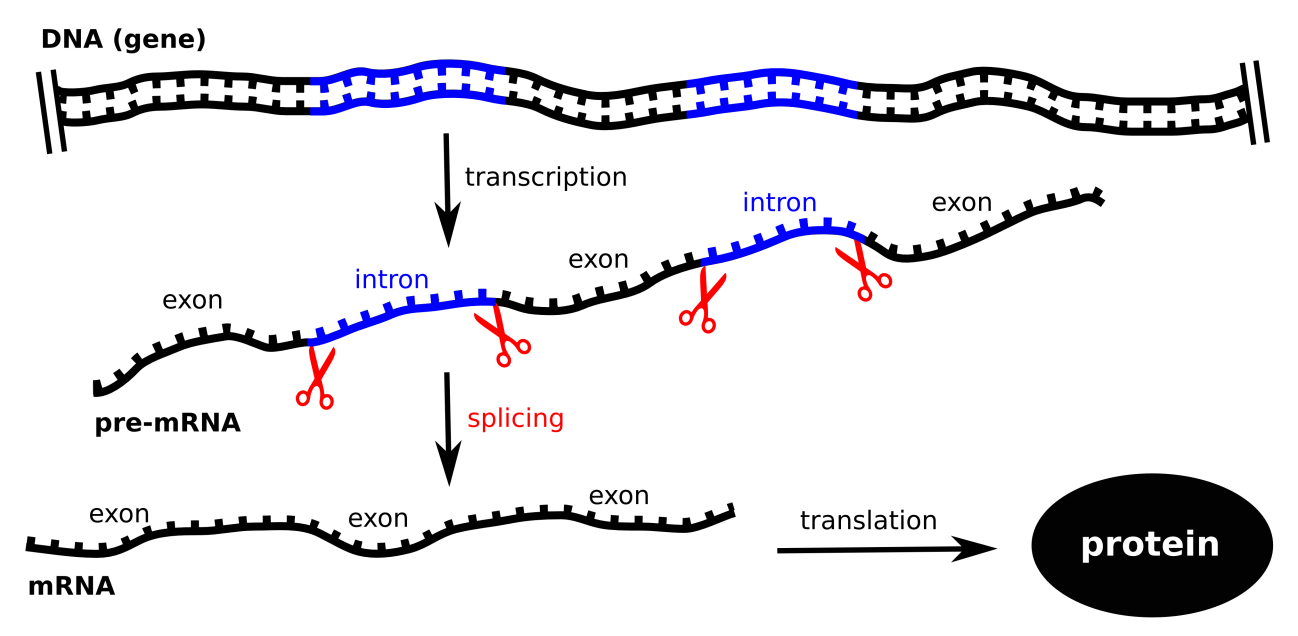


Dos Cromossomos às Proteínas: Transcrição



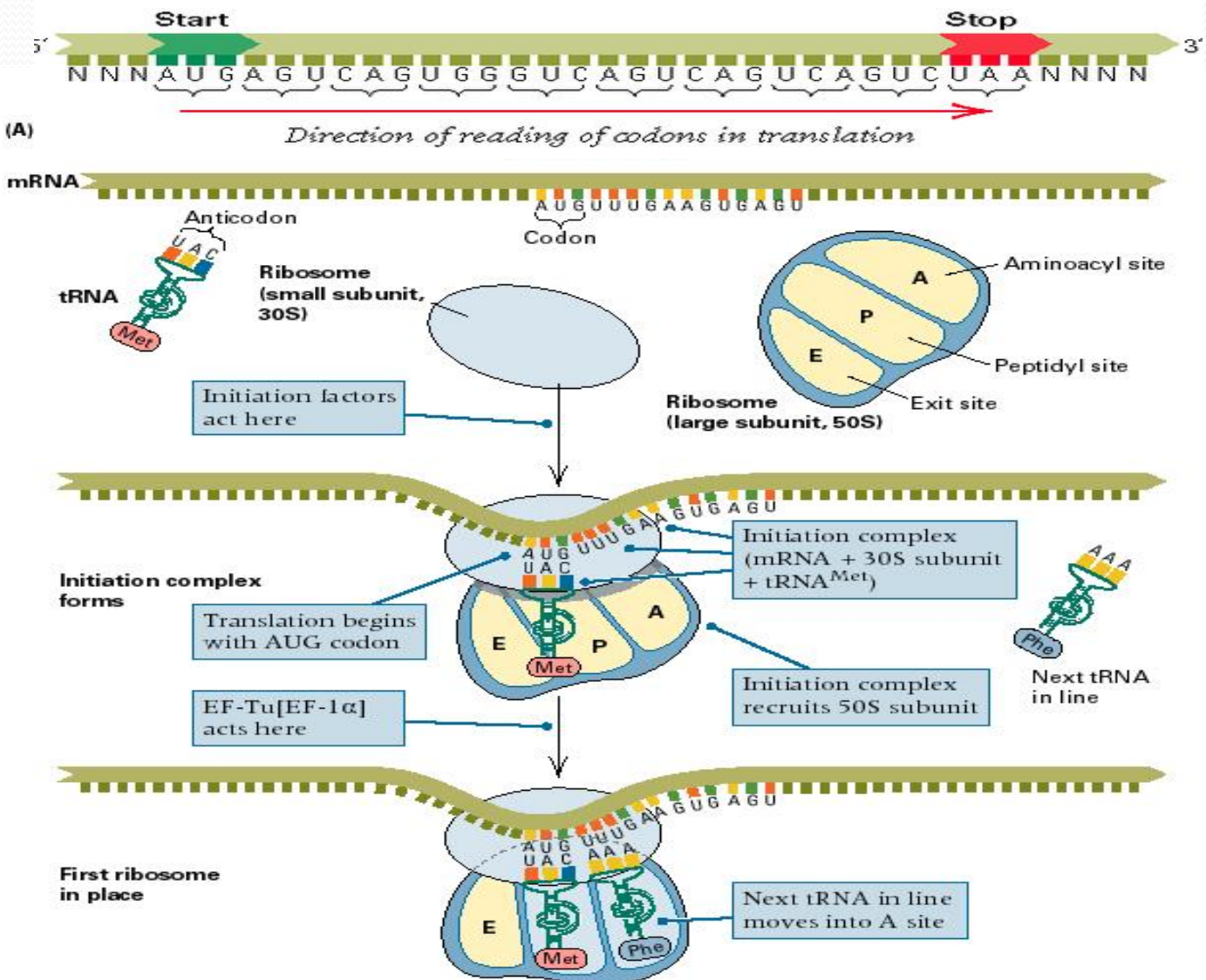
Fonte:http://geneed.nlm.nih.gov/topic_subtopic.php?tid=15&sid=22

Dos Cromossomos às Proteínas: Splincing

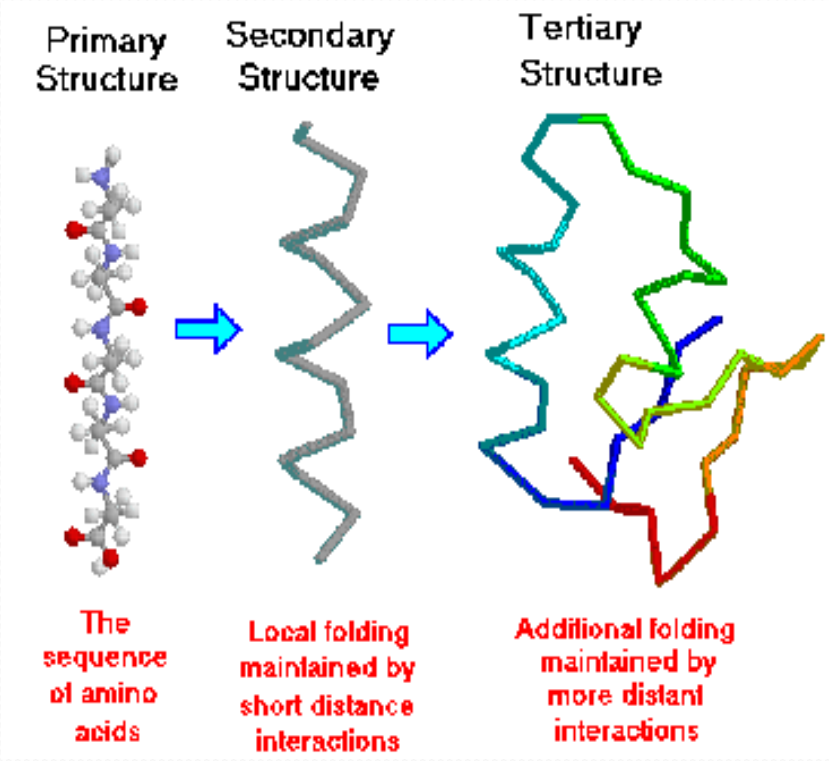


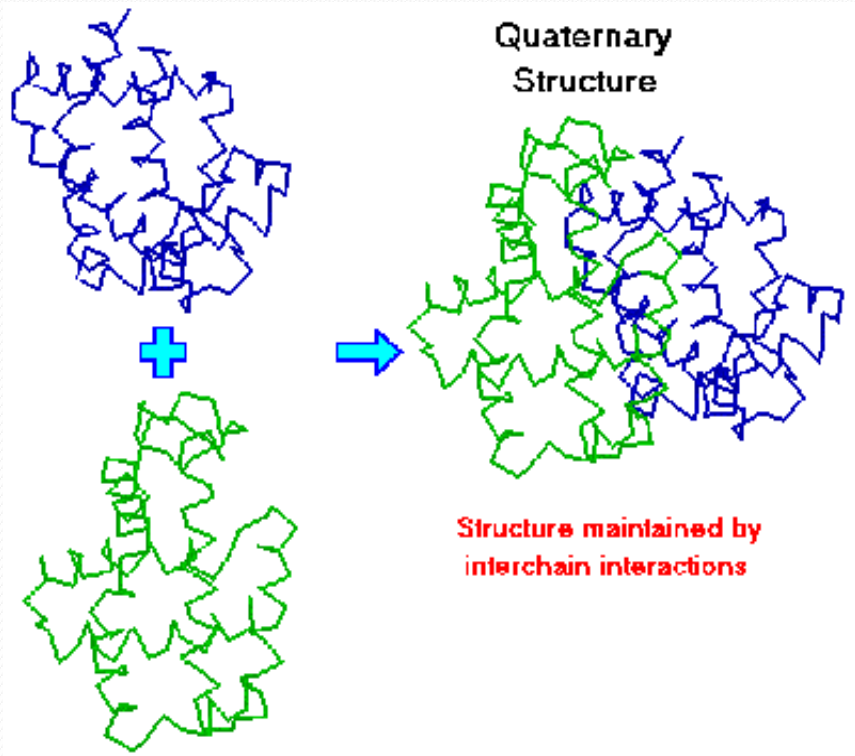
Fonte:http://www.nmrbordeaux.org/news/news.html

Tradução (mRNA - Proteina)

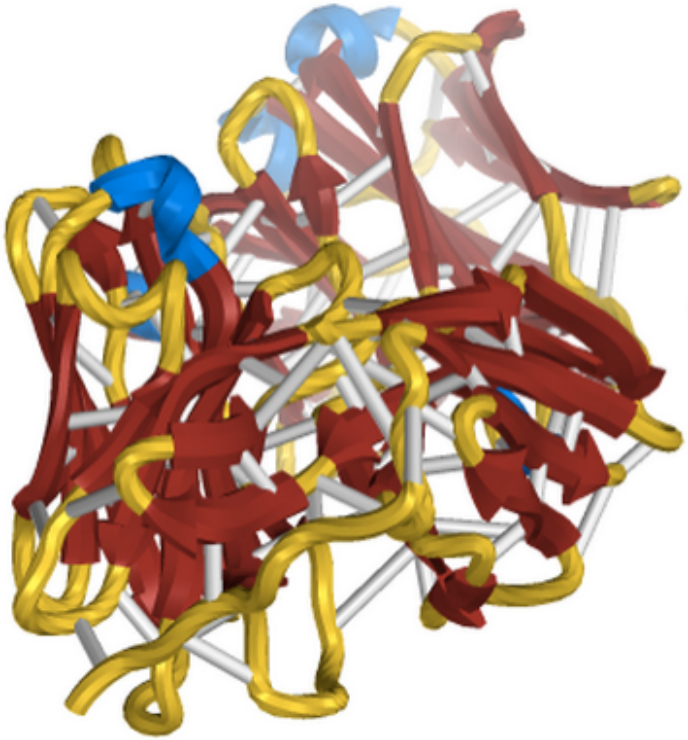


Proteinas – estrutura





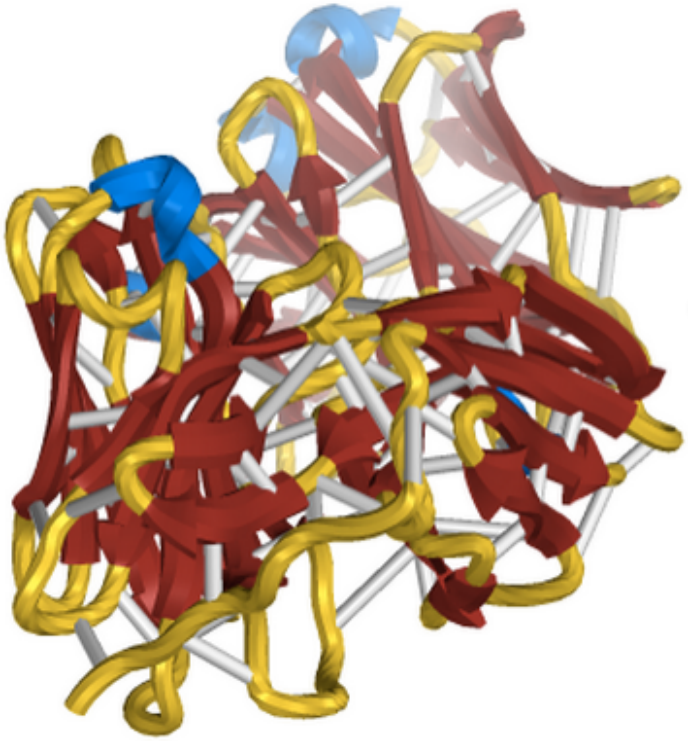
Proteínas – estrutura e função



Para biólogos

Fonte: http://3dprint.nih.gov/discover/3dpx-001795

Proteínas – estrutura e função



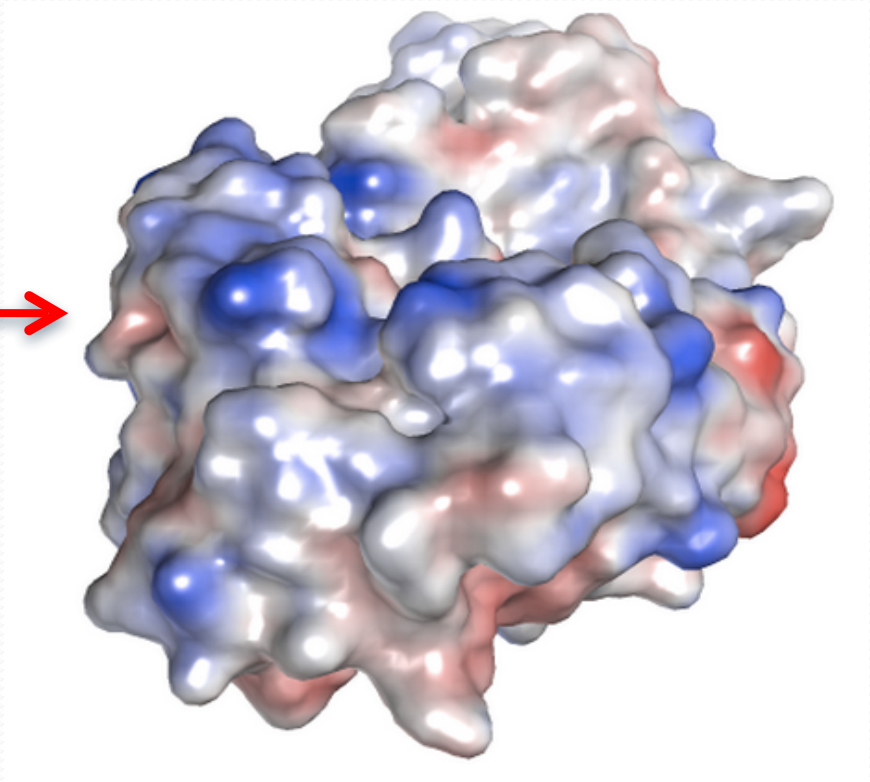
Alfa-hélices, beta folhas,
Sítios de ligação, etc.

Para biólogos

Fonte: http://3dprint.nih.gov/discover/3dpx-001795

Proteínas – estrutura e função

Para computeiros

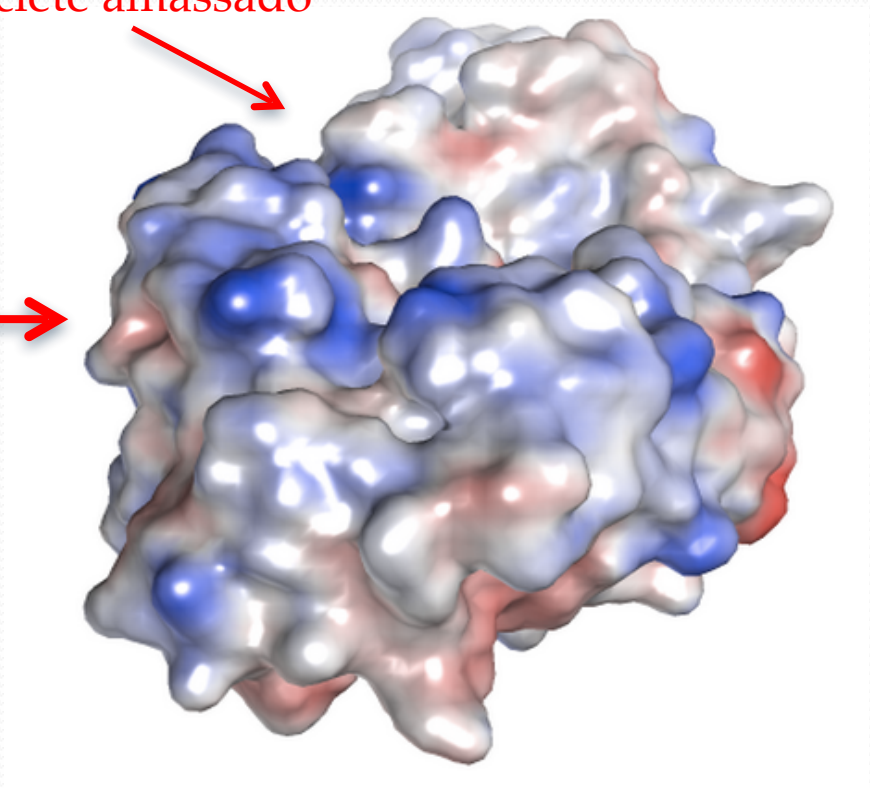


Fonte: http://3dprint.nih.gov/discover/3dpx-001795

Proteínas – estrutura e função

"chiclete amassado"

Para computeiros



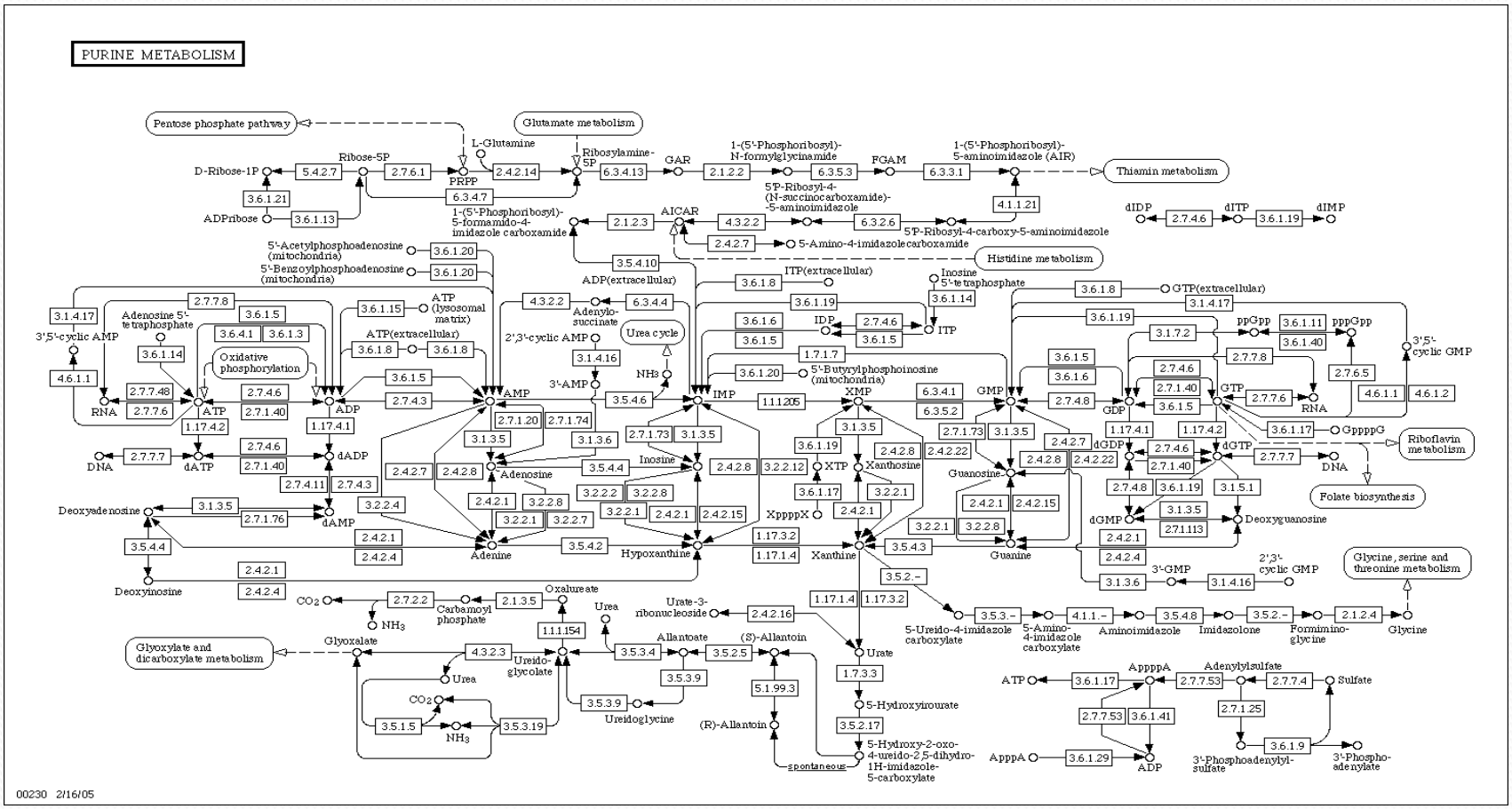
- Estrutura tridimensional importante para entender mecanismos de atuação da proteína
  - "relevo" e "cargas" explicam interações

Fonte: http://3dprint.nih.gov/discover/3dpx-001795

Redes Metabólicas

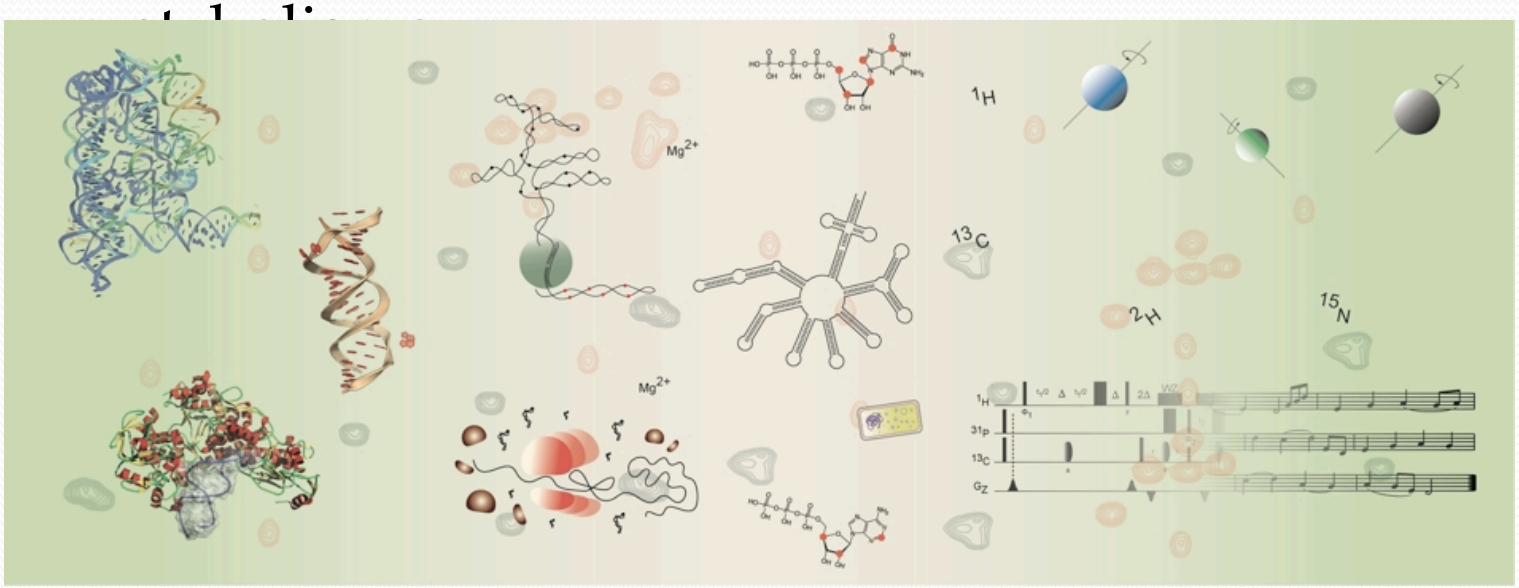
- Sistemas biológicos são o produto de interações complexas entre moléculas
- Estas moléculas formam redes complexas e estímulos e repressões
- O entendimento profundo do proceso celular vem do estudo destas redes

Redes Metabólicas: exemplo (met. purinas)



Não acabou: o “mundo”do RNA

- Nas últimas décadas descobrimos que apenas as proteínas não explicavam a complexidade dos seres
- A maior parte do genoma é transcrita e os RNAs não traduzidos formam componentes importantes de nosso



Fonte: http://www2.chem.umd.edu/groups/dayie/

Podemos olhar as áreas da Bionformática baseados no Dogma

- DNA e GENOMA:
  - **Sequenciamento**: descobrindo as bases

    eliminacao de contaminantes,, trimagem de primers e vetores, descartar sequências com baixa qualidade

    montagem de sequencias (varios pedaços, quebra cabeças)
- **mRNA e TRANSCRIPTOMA**
  - **Anotação**: onde estão os genes e o que fazem?

    predição de genes – RNA , Proteinas

    Caracterização de grupos de genes – função por semelhança

    Caracterização de RNAs não codificantes

Podemos olhar as áreas da Bionformática baseados no Dogma

- ESTRUTURA TRI-DIMENSIONAL
  predição de estrutura tridimensional
    tRNA, ncRNA
    Proteinas

- ENCONTRANDO REDES METABÓLICAS: biologia de sistemas
  análise de dados de expressão de genes
  diagnóstico
  descobrimento de vias metabólicas
  Simulação

- EVOLUÇÃO DO DNA: filogenia molecular
  relação evolutiva entre genes ou espécies
  evolução no sentido darwiniano, por favor.

De onde viemos: a Bioinformática no Brasil

- Brasil: AX vs DX
  - (Antes do Projeto Xylella vs. Depois do Projeto Xylella)
- Domínio da tecnologias modernas de seqüenciamento

  projetos genoma

  análise de dados de expressão (quais RNAS mensageiros foram transcritos e quando)

- não paramos aí
  - modelagem e dinâmica de proteínas
  - simulação de redes metabólicas
  - Biologia de sistemas
  - etc.

Qual a grande diferença?

- introdução do uso intensivo de ferramentas computacionais
  - métodos biológicos extremamente caros
  - projetos genoma propiciaram infinidade de informações e dados para análise
  - processos computacionais podem ser utilizados para análise exaustiva
- importância da validação
  - **computação é atividade *meio* não atividade *fim***
  - métodos computacionais e estatísticos são poderosos geradores de hipóteses
  - métodos biológicos, associados a análise estatística cuidadosa validam hipóteses
  - cuidado: análises nunca são melhores que a qualidade dos dados

Para onde vamos? (uma perspectiva de CC)

- análises mais rápidas: paralelização
- novas análises:desenvolvimento de novas ferramentas de análise
- construindo análises mais rápido: plataformas de análise

Processamento paralelo

- inúmeros problemas em bioinformática sofrem de limitações computacionais:
  - alinhamento múltiplo
  - busca de similaridade
  - estrutura de proteínas
  - predição de estruturas secundárias
  - etc.

Processamento paralelo

- paralelismo muitas vezes propagado como uma solução "mágica" para o problema.
- porém, podemos usar o termo paralelismo em vários sentidos
- muitas cpus vs muitos computadores
  - multiprogramação vs. algoritmos paralelos vs processamento distribuído

Processamento paralelo

- Aplicabilidade restrita de cada um.
  - multi-programação utilizada para executar programas que rodam independentemente e podem rodar ao mesmo tempo, seja dividindo apenas uma cpu ou várias
  - processamento distribuído indicado para realizar tarefas diferentes que cooperam entre si
  - algoritmos paralelos dividem uma tarefa individual em várias tarefas que podem ser executadas simultaneamente
- Algoritmos paralelos são a tecnologia mais promissora
  - Mais em evidência com arquiteturas multi-core e com novas tecnologias de integração
  - projetos de paralelização são complexos e nem sempre viáveis
- Uso de GPUs em Bioinformática parece particularmente promissor

Novas ferramentas

- uma das áreas mais desafiadoras da bioinformática
- uma das áreas mais complexas de pesquisa
  - necessidade de interação próxima de pesquisadores de áreas complemente distintas: ciências da vida vs. ciências "exatas"
  - metodologias diferentes
  - linguagens diferentes
  - cuidados formais diferentes
  - visão diferente da questão de sigilo
  - visão diferente do que são resultados satisfatórios
- porém é a área onde os grandes saltos de qualidade podem ser alcançados
  - afirmação inclui novas ferramentas não computacionais

Novas ferramentas de bioinformática (continuação)

- Simuladores
  - modelos matemáticos e computacionais para reprodução de comportamentos interdependentes
  - simuladores paralelos de redes metabólicas
  - sistemas dinâmicos com equações diferenciais
  - ....
- Análise combinatória
  - alinhamento múltiplo e análise filogenética
  - detecção automática de inversões

Novas ferramentas (continuação.)

- Otimização e Análise Numérica
  - redução de dimensionalidade (microarrays)
  - sistemas dinâmicos
- Bancos de dados
- Aprendizado Computacional e reconhecimento de padrões
- Biologia de sistemas

Novas ferramentas de bioinformática : bancos de dados

- realidade
  - Volume extremamente grande de novos dados (apenas o Genoma Humano tem 4 bilhões de pares de bases, ao menos 30.000 genes com até 100.000 variantes, sem falar de todas as possíveis redes metabólicas)
  - Até projetos de sequenciamento de porte médio envolvem milhares milhões de sequências
  - Número inimaginável de possíveis relações
  - Grande parte dos dados ainda guardados em arquivos simples não estruturados (.txt, .doc, .xls, ...)
  - Área extremamente promissora

Novas ferramentas: reconhecimento de Padrões

- criação de função matemática para caracterização de fenômenos
- isso me interessa?
  - Encontrar os genes de um novo genoma
  - Quais regiões determinam a ativação de um gene?
  - Como descobrir RNAs não codificastes funcionais -
  - famílias de proteínas/genes
  - processamento de imagens (reconhecimento de características, limpeza de ruído, etc.)
- algumas tecnologias
  - gramáticas estocásticas, modelos de covariancia
  - HMMs – Cadeias Ocultas de Markov
  - redes neurais
  - morfologia matemática

Aprendizado Computacional

- diferente de sistemas inteligentes
- utilizado quando conhecimento biológico não é suficiente
- Aprendizado supervisionado:
  - Conhecemos dados em categorias diferentes e queremos criar classificador:
    - Diagnóstico de câncer baseado em dados de expressão de genes
    - Caracerizaçãod eu ma família de genes
  - **necessário um bom conjunto de caracterização**
- Aprendizado não supervisionado (clusterização)
  - Utilizamos medidas para agrupar elementos de um universo em conjuntos

Um exemplo: gramáticas

- Gramáticas: mecanismo de descrição de uma linguagem
- Linguagem: um conjunto de frases
  - nucleotídeos são as palavras
  - seqüências são as frases

ou

  - codons são as palavras
  - proteínas são as frases
- Gramáticas podem ser utilizadas para
  - Reconhecimento
  - Geração
  - Aprendizado computacional

Um exemplo familiar

- Frase ::= sujeito predicado
- sujeito ::= artigo nome
- artigo ::= a | o
- nome ::= cão | moça | dia
- predicado ::= verbo adjetivo
- verbo ::= está | estava
- adjetivo ::= feliz | triste

Construindo uma frase

Frase ->

Construindo uma frase

Frase ->

->sujeito predicado ->

Construindo uma frase

Frase ->

->sujeito predicado ->

-> artigo nome predicado ->

Construindo uma frase

Frase ->

->sujeito predicado ->

-> artigo nome predicado ->

-> a nome predicado

Construindo uma frase

Frase ->

->sujeito predicado ->

-> artigo nome predicado ->

-> a nome predicado

-> a moça predicado ->

Construindo uma frase

Frase ->

->sujeito predicado ->

-> artigo nome predicado ->

-> a nome predicado

-> a moça predicado ->

->a moça verbo adjetivo->

Construindo uma frase

Frase ->

->sujeito predicado ->

-> artigo nome predicado ->

-> a nome predicado

-> a moça predicado ->

->a moça verbo adjetivo->

-> a moça está adjetivo ->

Construindo uma frase

Frase ->

->sujeito predicado ->

-> artigo nome predicado ->

-> a nome predicado

-> a moça predicado ->

->a moça verbo adjetivo->

-> a moça está adjetivo ->

-> a moça está feliz

Construindo uma frase

Frase ->

->sujeito predicado ->

-> artigo nome predicado ->

-> a nome predicado

-> a moça predicado ->

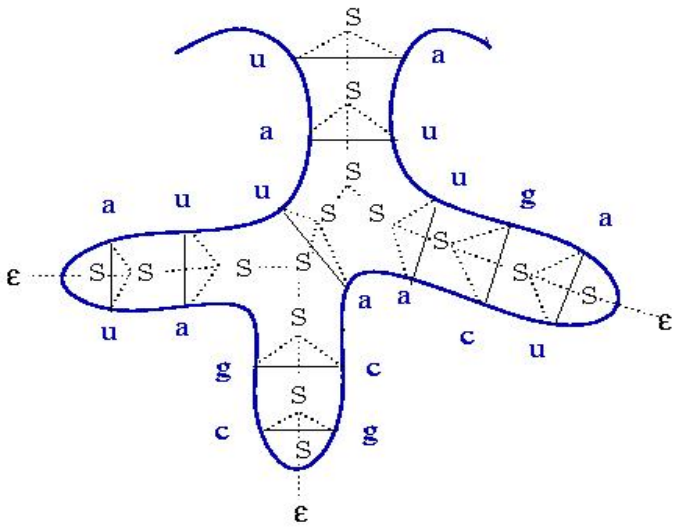
->a moça verbo adjetivo->

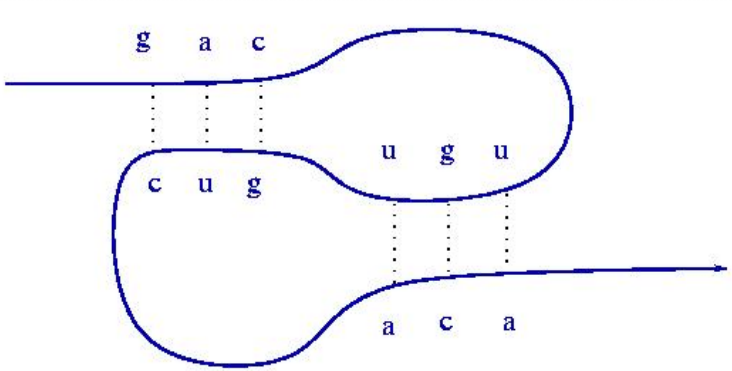
-> a moça está adjetivo ->

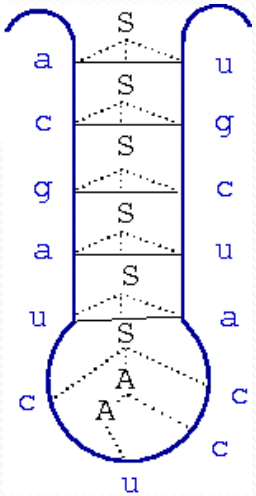
-> a moça está feliz

-> a moça está feliz

Exemplos de fenômenos genéticos que podem ser modelados







Gramática livre de contexto

- S ::= a S u | u S a | c S g | g S c | A
- A ::= A a | A u | A c | Ag | nil

S -> aSu->

->acSgu->

->acgScgu->

->acgaSucgu->

->acgauSaucgu ->

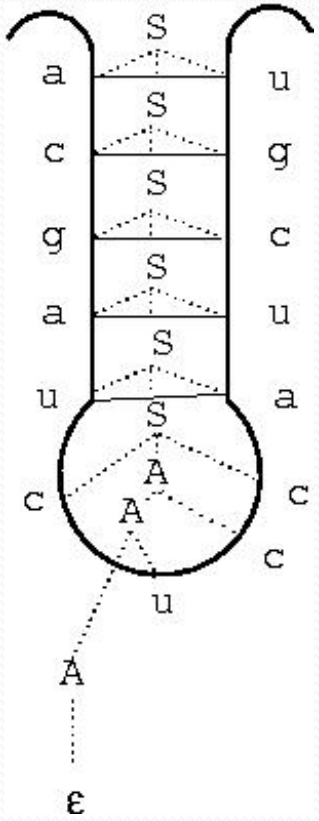
->acgauAucgu->

->acgauAcucgu->

->acgauAccucgu->

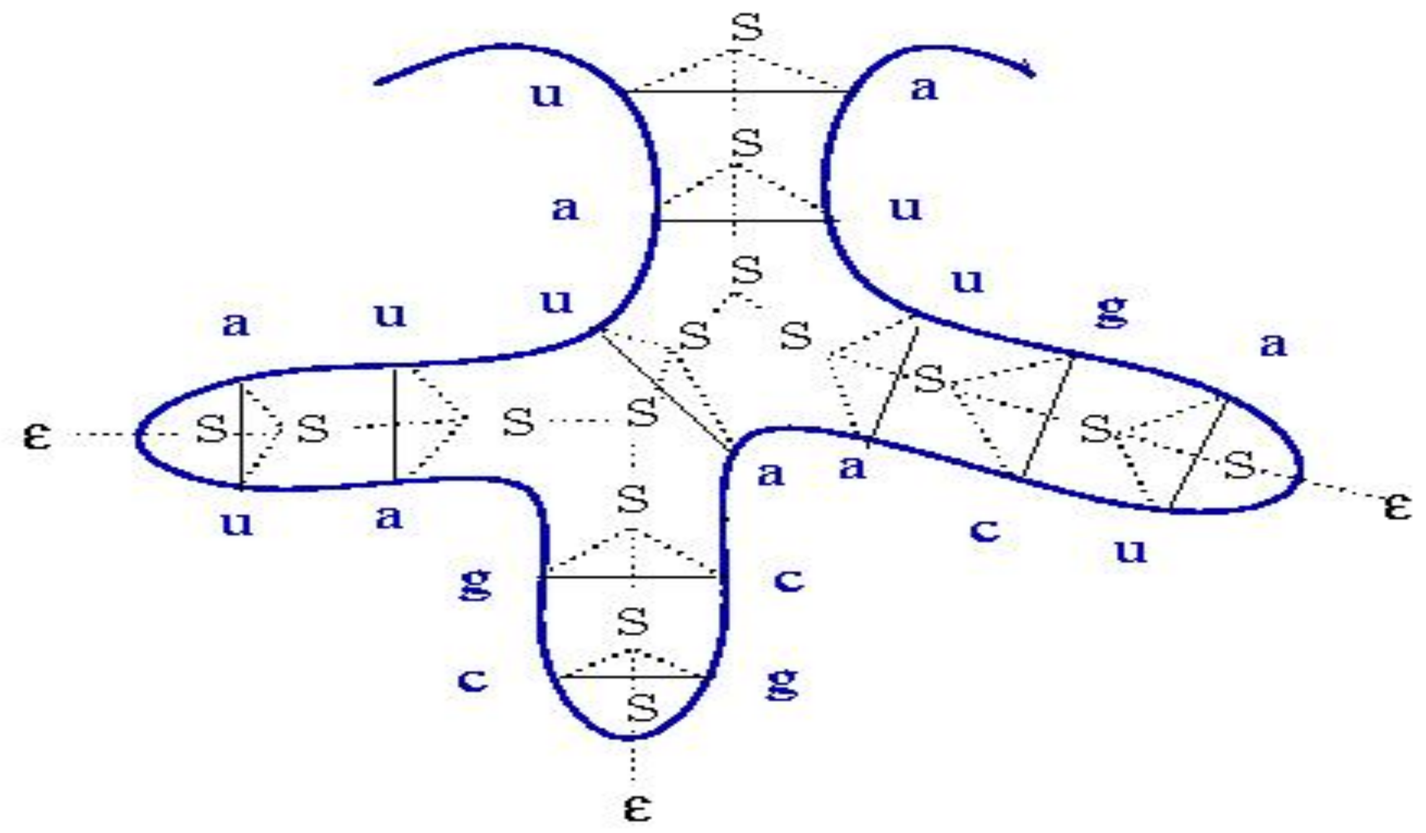
->acgauAuccugu->

->acgauuccugu



Gramática livre de contexto

- S ::= a S u | u S a | c S g | g S c | SS



Desenvolvimento de software

- desenvolvimento moderno de software
  - modularidade
  - abstração
  - encapsulamento
- encapsulamento
  - detalhes de implementação do software não precisam ser conhecidos nem pelos usuários, nem pelos sistemas que os utilizam
  - visão dos componentes é abstrata, apresentando um modelo em geral mais simples da realidade

Desenvolvimento de software: abstração

- modularidade
  - desenvolvimento de componentes facilmente intercambiáveis
  - ex. trocar dois programas de alinhamento sem nenhum trabalho adicional
  - interface padrão
  - e.g. motor de automóvel, caixa de câmbio,...
- plataformas de análise
  - criação de ambientes para desenvolvimento de sistemas específicos
  - exemplo similar: programas de desenho

    paintbrush

    corel draw

    autocad

Plataformas de análise

- aumento grande de produtividade para desenvolvimento de novos sistemas
  - usuário compõe soluções a partir de elementos pré definidos (triangulos, retangulos, retas)
- baixo custo de customização para necessidades específicas
  - componentes configuráveis (mudança de cor, de traço na linha, espessura de linha)
- gostaríamos de algo semelhante para processamento de dados biológicos
- poucos sistemas até agora na área de bioinformática
- amplo espectro de desenvolvimento

Desenvolvimento de plataformas: questões a tratar

- inicialmente definir escopo de atuação
- análise das soluções atuais
- existe boa definição das tarefas?
- soluções apresentam estrutura comum?
- podemos isolar componentes de soluções que se repetem?
- criar um modelo para integração dos componentes para, a partir deles, descrever as soluções atuais

Desenvolvimento de Plataformas: ToPS

-

Lições a lembrar

- interação interdisciplinar é essencial
- o pesquisador de C. Computação
  - modelagem computacional flexível e fácilmente extensível
  - entendimento ao menos parcial do problema para criação de um modelo computacional que registre as características do problema
- o pesquisador da área do domínio do problema
  - ententimento da variabilidade das soluções
  - auxílio no desenvolvimento da interface de configuração de soluções: a ferramenta precisa falar “biologuês”.
  - análise critica dos resultados

Uma lição importante: uma andorinha só não faz verão

- pode ser um erro grave:
  - **uso de ferramentas computacionais sem a devida compreensão de suas limitações**

    blast vs s.waterman vs. n. wunch

    alinhamento múltiplo e análise filogenética
  - **uso de técnica computacional sem a compreensão da da biologia**

    análise filogenética e hipóteses de evolução

    psi-blast, matrizes de substituição

    predição de genes

    ...

Conclusões

- área de bioinformática tem amplo espectro
- conhecimentos nas duas áreas é importante
  - interação entre pesquisadores
  - formação multi-disciplinar
- várias abordagens
  - desenvolvimento de sistemas e/ou plataformas de análise
  - bancos de dados
  - uso de novas tecnologias
    aprendizado computacional
  - ...

Conclusões

- Alguns temas não abordados
  - genética de populações
  - análise filogenética
  - modelagem e dinâmica de proteínas
  - construção automática de redes metabólicas
  - análise de expressão gênica
  - Processamento de dados médicos
  - Neurociência
- Algumas áreas não citadas mas ativas na área
  - Teoria dos grafos e Combinatória
  - Processamento de Imagens
  - Sistemas Dinâmicos

Como estudar bionformática

- Nosso departamento é um dos centros importantes de pesquisa e ensino em bionformática
  - Sede do primeiro programa de doutorado na área no Brasil (2002)
- Bionformática é parte da trilha de E-science, mas é possível atuar a partir de outras áreas
- Iniciação científica com um dos vários professores atuantes na área
  - Alair P. Do Lago, Carlos Eduardo Ferreira, **Alan M. Durham**, André Fujita, João Eduardo Ferreira, Kelly Braguetto, Roberto Hirata, Ronaldo Hashimoto

OBRIGADO!